SiT: Exploring Flow and Diffusion-based Generative Models with Scalable Interpolant Transformers

Abstract

We present Scalable Interpolant Transformers (SiT), a family of generative models built on the backbone of Diffusion Transformers (DiT). The interpolant framework, which allows for connecting two distributions in a more flexible way than standard diffusion models, makes possible a modular study of various design choices impacting generative models built on dynamical transport: learning in discrete or continuous time, the objective function, the interpolant that connects the distributions, and deterministic or stochastic sampling. By carefully introducing the above ingredients, SiT surpasses DiT uniformly across model sizes on the conditional ImageNet 256x256 and 512x512 benchmark using the exact same model structure, number of parameters, and GFLOPs. By exploring various diffusion coefficients, which can be tuned separately from learning, SiT achieves an FID-50K score of 2.06 and 2.62, respectively.